MISSING DATA IMPUTATION USING COMPRESSIVE SENSING TECHNIQUES FOR CONNECTED DIGIT RECOGNITION

Jort Gemmeke and Bert Cranen

Centre for Language and Speech Technology
Radboud University Nijmegen
The Netherlands

ABSTRACT

An effective way to increase the noise robustness of automatic speech recognition is to label noisy speech features as either reliable or unreliable (missing) prior to decoding, and to replace the missing ones by clean speech estimates. We present a novel method based on techniques from the field of Compressive Sensing to obtain these clean speech estimates. Unlike previous imputation frameworks which work on a frame-by-frame basis, our method focuses on exploiting information from a large time-context. Using a sliding window approach, denoised speech representations are constructed using a sparse representation of the reliable features in an overcomplete dictionary of clean, fixed-length speech exemplars. We demonstrate the potential of our approach with experiments on the AURORA-2 connected digit database.

Index Terms— Compressive Sensing, Missing Data Techniques, Noise robustness, ASR

1. INTRODUCTION

Automatic speech recognition (ASR) performance degrades substantially when speech is corrupted by background noise not seen during training. Fortunately, this effect can be mitigated by Missing Data Techniques (MDTs) both for stationary and non-stationary noises and for a wide range of Signal-to-Noise (SNR) ratios (e.g. [1]). At the heart of MDT is the assumption that it is possible to estimate −prior to decoding− which spectro-temporal elements of the acoustic representations are reliable (i.e., dominated by speech) and which are unreliable (i.e., dominated by background noise). During decoding these reliability estimates, referred to as *spectrographic masks*, can then be used to favor reliable features over unreliable ones.

Two different missing data approaches exist to handle unreliable data, viz. *marginalization* and *imputation*. With marginalization [2] missing values are largely ignored during the decoding by integrating over their possible ranges. With imputation [3, 4, 5, 6], missing features are replaced by clean speech estimates. In this paper we will focus on imputation.

Although previously developed *missing data imputation* methods have achieved impressive gains in recognition accuracy, a ceiling effect occurs at SNRs $\leq$ 0 dB. At these low SNRs the recognition performance is insufficient for practical applications. This is most likely due to the fact that these methods work on a frame-by-frame basis (i.e. strictly local in time): At SNRs $\leq$ 0 dB a substantial number of frames may contain few, if any reliable features and clearly such frames are unlikely to contain sufficient information for successful imputation.

This data scarcity could be avoided if one were able to exploit the coherence over time and frequency that speech signals exhibit. By taking more time-context into account reliable features from neighboring frames could also be utilized, thus taking advantage of the inherent redundancy of speech signals. However, with conventional MDTs (e.g. [3]) which are based on parametric models, incorporating more time context would quickly yield unwieldy procedures involving huge numbers of (co)variance parameters to be estimated.

In [7], we introduced an alternative, non-parametric approach to imputation based on methods from the emergent field of *Compressive Sensing (CS)*: We dubbed it *sparse imputation (SI)*. In essence it is an exemplar-based method which, prior to decoding, replaces unreliable ('missing') features by clean speech estimates. It can therefore be used as a front-end for an arbitrary ASR engine. The data scarcity problem associated with single frame approaches is tackled by using exemplars that span multiple frames.

The CS theory asserts that if a signal is sparse or compressible in some dictionary, it can be recovered using a very limited number of measurements. In [8] it was shown that the CS framework can be used for missing data imputation. By treating the reliable features as the only available measurements from which the underlying unknown sparse signal must be obtained, the missing spectro-temporal elements of the signal can be reconstructed simply by projecting the sparse signal in the dictionary. Because the dictionary contains complete vectors from which no elements are missing, the projection operation yields complete vectors as well, thereby effectively imputing any missing data.

In [9] it was suggested that a signal might be very sparsely represented in an overcomplete dictionary of *exemplars* by expressing that signal as a linear combination of a small number of example signals. In our sparse imputation method we follow this approach, representing speech tokens as a linear combination of tokens from an overcomplete dictionary of of noise-free exemplars represented by fixed length vectors (in [7] speech tokens and exemplars were chosen to constitute whole words). For an unknown speech token, a sparse linear combination is sought in this dictionary using all reliable features of the *entire* token. The weights of the linear combination are then used to compute the projection in this dictionary to provide clean speech estimates with which the unreliable features must be replaced.

There is a substantial amount of work on source separation using sparse, possibly overcomplete, representations (e.g. [10, 11, 12]).

However, in contrast to our SI approach, these methods invariably decompose the signal using models trained on the individual sources. Due to the great variety of possible background noises, it is virtually impossible to find a generic noise source model. In our work we therefore try to avoid making assumptions about the corrupting noise by using a missing data mask which focusses on speech properties only.

For a single digit recognition task and whole word exemplars it was shown in [7] that the Sparse Imputation method is able to restore the underlying clean speech even in very low SNR conditions, provided a sufficiently accurate spectrographic mask can be created. Using whole digit speech tokens limits the applicability of the method to situations where the word boundaries are known beforehand (i.e., to isolated word recognition). The goal of this paper is to investigate whether it is feasible to extend the SI framework to allow for imputation of utterances in which word-boundaries are *not* known in advance (i.e. to continuous speech recognition). As a first step towards a continuous speech recognition task we explore its performance on the AURORA-2 connected digit recognition task [13]. In order to assess the strengths and weaknesses of the proposed approach, we compare two sets of recognition accuracy results. The first set is obtained from a state-of-the-art recognizer which uses the spectrographic mask for a frame-based, state dependent imputation in its back-end [14]. The second set of results is obtained with the same recognizer, fed with cleaned speech estimates provided by our sparse imputation front-end (in combination with a mask that considers all features reliable).

Because the performance of an MDT recognizer is known to be dependent on the quality of the used spectrographic mask, we consider recognition accuracy for two types of masks: (1) an 'oracle' mask and (2) an estimated, harmonicity mask which has proved to give good results when used with the back-end imputation method in [14]. We compare recognition accuracies for both types of masks and for both types of imputation and study how recognition performance depends on the overlap between the sliding imputation windows.

Sparse Imputation introduces several new parameters, such as the number of clean exemplars and the duration (number of frames) of these exemplars. For continuous speech at least one additional parameter is introduced, viz. the step size with which exemplars are matched against the signal to be processed. Although the additional parameters might well show significant interactions, this paper will only investigate these parameters in isolation.

The rest of the paper is organized as follows. In Section 2 we briefly describe MDT. In Section 3 we introduce the sparse imputation framework for word-like units. In Section 4 we extend this framework for use in continuous ASR. In Section 5 we compare recognition accuracies with the baseline decoder and we give our conclusions in Section 7. Finally, we discuss future work in Section 8.

2. MISSING DATA TECHNIQUES

In ASR, speech representations are typically based on some spectro-temporal distribution of acoustic power, called a spectrogram. In noise-free conditions, the value of each element in this two-dimensional matrix is determined by the speech signal only. In noisy conditions, however, the acoustic power in each cell may (in part) be due to background noise. Under the assumption that the noise is additive and uncorrelated to the speech, the power spectrogram of noisy speech $\boldsymbol{Y}$, can be described as the sum of the individual spectrograms of clean speech $\boldsymbol{S}$ and noise $\boldsymbol{N}$, i.e., $\boldsymbol{Y} = \boldsymbol{S} + \boldsymbol{N}$.

Elements of $\boldsymbol{Y}$ that predominantly contain speech energy are distinguished from those dominated by noise energy by introducing a spectrographic mask. With all spectrograms represented as $K \times T$ dimensional matrices (K being the number of frequency bands and T the number of time frames), a mask is defined as an equally sized matrix in which elements with value 1 indicate that the corresponding cell of $\boldsymbol{Y}$ is dominated by speech ('reliable'), while 0 means that it is dominated by noise ('unreliable' c.q. 'missing'). Thus, we write:

$$M(k,t) = \begin{cases} 1 \stackrel{def}{=} \text{reliable} & \text{if } S(k,t)/N(k,t) > \theta \\ 0 \stackrel{def}{=} \text{unreliable} & \text{otherwise} \end{cases} \quad (1)$$

with some empirical threshold θ, frequency band k ($1 \leq k \leq K$) and time frame t ($1 \leq t \leq T$). Representing the power spectrum of the noisy speech on a log-compressed scale, reliable features Y_r may be written as:

$$\begin{aligned} \log[Y_r(k,t)] &= \log[S(k,t) \cdot (1 + N(k,t)/S(k,t))] \\ &\approx \log[S(k,t)] \end{aligned} \quad (2)$$

In other words, under the assumption of additive background noise, reliable noisy speech features can be used directly as estimates of their clean speech feature counterparts.

In experiments with artificially added noise, the mask can be computed using knowledge about the corrupting noise and the clean speech signal, the so-called *oracle masks*. In realistic situations, however, the masks must be estimated. Many different mask estimation techniques have been proposed; see [15] and the references therein for a comprehensive overview. These include techniques such as SNR based estimators [16], methods that focus on speech characteristics, e.g. harmonicity based SNR estimation [14] and mask estimation by means of Bayesian classifiers [17]. In Section 5 we will use the harmonicity mask [14] to investigate the properties of our method in combination with an estimated mask.

The SI method proposed in Section 3 is an example of a *feature vector imputation* method [1]. Because it is applied prior to decoding, it may be viewed as a front-end data cleaning technique which can be used in combination with any conventional ASR system. Recognition can then be performed as if all features were reliable. Imputation techniques, however, can also be integrated in the back-end of an ASR engine as illustrated by a successful approach called *class conditional imputation (CCI)* [1, 6]. The latter approach, which we will use in Section 5 as a baseline to compare our new method with, makes the clean speech estimates dependent on the hypothesized state of the hidden Markov model.

3. WORD BASED IMPUTATION

3.1. Sparse representation of word-like units

We express the $K \times T$ log-compressed power spectrogram matrix (cf. Eq. 2) of a word-like unit $\boldsymbol{S}$ as a single vector $\boldsymbol{s}$ of dimension $D = K \cdot T$ by concatenating T subsequent time frames. We assume T to be fixed. As in [7], we consider the noisy speech $\boldsymbol{y}$ as a linear combination of exemplar spectrograms $\boldsymbol{a}_n$, where n, ($1 \leq n \leq N$) denotes a specific exemplar in the set of N available exemplars. We

write:

$$\boldsymbol{s} = \sum_{n=1}^{N_A} x_n \boldsymbol{a_n} = \boldsymbol{A}\boldsymbol{x} \tag{3}$$

with $\boldsymbol{A} = (\boldsymbol{a_1}\ \boldsymbol{a_2} \ldots \boldsymbol{a_{N-1}}\ \boldsymbol{a_N})$ a matrix with dimensionality $D \times N$ and $\boldsymbol{x}$ an N-dimensional weight vector.[1]

Typically, the number of exemplar spectrograms will be much larger than the dimensionality of the acoustic representation ($N \gg D$). Therefore, the system of linear equations (3) has no unique solution. Research in the field of *compressive sensing* [18, 19] has shown however that if $\boldsymbol{x}$ is *sparse*, $\boldsymbol{x}$ can be determined *uniquely* by solving:

$$\min_{\boldsymbol{x}} \{ \|\boldsymbol{x}\|_0 \} \text{ subject to } \boldsymbol{s} = \boldsymbol{A}\boldsymbol{x} \tag{4}$$

with $\|.\|_0$ the l^0 zero norm (i.e., the number of nonzero elements).

3.2. l^1 minimization

The combinatorial problem in Eq. 4 is NP-hard and therefore cannot be solved in practical applications. However, it has been proven that, with mild conditions on the sparsity of $\boldsymbol{x}$ and the structure of $\boldsymbol{A}$, $\boldsymbol{x}$ can be determined [20] by solving:

$$\min_{\boldsymbol{x}} \{ \|\boldsymbol{x}\|_1 \} \text{ subject to } \boldsymbol{s} = \boldsymbol{A}\boldsymbol{x} \tag{5}$$

This convex minimization problem can be cast as a least squares problem with an l^1 penalty:

$$\min_{\boldsymbol{x}} \{ \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{s}\|_2 + \lambda \|\boldsymbol{x}\|_1 \} \tag{6}$$

with a regularization parameter λ. If $\boldsymbol{x}$, with sparsity $f = \|\boldsymbol{x}\|_0$, is very sparse, Eq. 6 can be solved efficiently in $\mathcal{O}(f^3 + N)$ time using homotopy methods [21].

3.3. Sparse imputation

When using noisy features, the reliable components of $\boldsymbol{y}$ serve as an approximation of the unknown clean speech features $\boldsymbol{s}$ (cf. Eq. 2). In order to find a solution of Eq. 6 which is only based on reliable features, we carry out a *weighted* norm minimization instead:

$$\min_{\boldsymbol{x}} \{ \|\boldsymbol{W}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{W}\boldsymbol{y}\|_2 + \lambda \|\boldsymbol{x}\|_1 \} \tag{7}$$

with $\boldsymbol{W}$ a diagonal matrix of weight coefficients. Although it is possible to allow the weights to assume any value in the range between 0 and 1 [22], in this paper we have opted for a simpler approach in which the weights are determined directly by the binary missing data mask $\boldsymbol{M}$ being either 0 or 1. By concatenating subsequent time frames of $\boldsymbol{M}$, similarly as we did for the spectrogram $\boldsymbol{Y}$, we construct a vector $\boldsymbol{m}$ to represent the weights on the diagonal of $\boldsymbol{W}$: $\text{diag}(\boldsymbol{W}) = \boldsymbol{m}$. Thus, we effectively use $\boldsymbol{W}$ as a row selector, picking only those rows of $\boldsymbol{A}$ and $\boldsymbol{y}$ that are assumed to contain reliable data.

As suggested in [8] it is possible to use the sparse representation $\boldsymbol{x}$ obtained from solving Eq. 7 to estimate the missing values of $\boldsymbol{y}$ by reconstruction:

$$\hat{\boldsymbol{y}} = \boldsymbol{A}\boldsymbol{x} \tag{8}$$

[1] We do not require that $\boldsymbol{x}$ is non-negative. Because $\boldsymbol{s}$ is expressed in the log domain, it might take negative values. We have, however, chosen the 0 dB level such that we hardly observe any negative values.

$\hat{\boldsymbol{y}}$ is obtained by a linear combination of corresponding elements of the dictionary vectors, the weights of which were determined using only reliable data. Hence, a version of $\hat{\boldsymbol{y}}$ that is reshaped into a $K \times T$ matrix can be considered a denoised spectrogram of the underlying speech signal.

3.4. Theoretical bounds on successful imputation

Obviously, no restoration is possible if $\boldsymbol{y}$ does not contain any reliable coefficients at all. In practice, a minimum number of reliable coefficients will be required for successful restoration of $\boldsymbol{y}$. If the reliable features were randomly distributed over the time-frequency plane we could consider the using of the missing data mask as randomly projecting the features to a lower dimensional subspace. Under such conditions reconstruction will be possible provided the number of reliable features does sufficiently exceed the sparsity of the sparse representation $\boldsymbol{x}$. In practice, however, the reliable features are not randomly distributed over the time-frequency plane but localized in coherent regions. Thus, successful reconstruction heavily depends on the structure of the matrix $\boldsymbol{WA}$.

While theoretical guarantees for successful recovery of $\boldsymbol{x}$ given a matrix $\boldsymbol{WA}$ do exist (cf. [18, 19, 8]) these are not of great practical value. First, these bounds, such as the Restricted Isometry Property (RIP), are sufficient but not strictly necessary conditions. As a result these bounds may represent rather pessimistic estimates; moreover, they are NP-hard to establish. Second, due to the fact that the spectrographic mask is both dependent on the individual utterance and the environmental noise condition, the dictionary matrix $\boldsymbol{WA}$ used for finding the sparse representation is also data dependent. This makes it unfeasible to come up with an estimate for the minimum number of reliable coefficients which is both practically useful and generally valid. Admittedly, it would be highly desirable to have a theoretically valid criterion for deciding when sparse imputation results become meaningless. By the lack thereof, however, we perform sparse imputation unconditionally, thus accepting the risk of a flawed restoration.

4. IMPUTATION OF CONTINUOUS SPEECH

The approach described in Sec. 3 is suitable for imputation of noisy speech tokens that can be adequately represented by a fixed number of time frames T [7]. Since arbitrary length utterances clearly do not satisfy this constraint, it cannot be applied to continuous speech recognition. In this section we extend the sparse imputation framework for use with speech signals of arbitrary length by using a sliding, fixed-length time window. Robustness against windows with few or no reliable features is provided by using overlapping windows.

4.1. Imputation using overlapping windows

We divide an utterance $\boldsymbol{Y}$ of T frames in a series of overlapping time windows of R frames and perform imputation for every individual window with the method described in Section 3. To this end we reshape $\boldsymbol{Y}$, as before, to a single vector $\boldsymbol{y}$ by concatenating subsequent time frames. With the dimensions of the spectrogram $\boldsymbol{Y}$ being $K \times R$, the windowed vectors derived from $\boldsymbol{y}$ have size $L = K \cdot R$ (cf. Fig. 1). Similarly, the dictionary $\boldsymbol{A}$ is formed by N exemplar

vectors, which are reshaped versions of spectrograms (each spanning also R frames). In contrast to the word based approach, however, the exemplar spectrograms in the dictionary $\boldsymbol{A}$ are now created by extracting spectrogram fragments with a random offset from randomly selected utterances in the clean train set. The dimensions of $\boldsymbol{A}$ are $L \times N$.

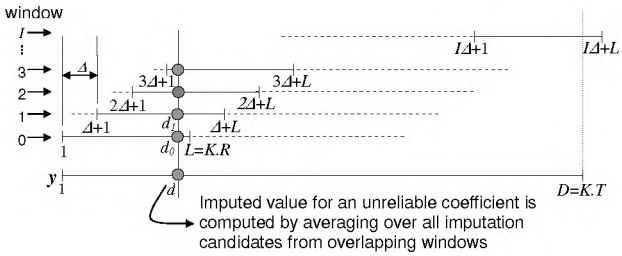

Fig. 1: Diagram of imputation using overlapping windows.

The number of windows needed for processing the entire speech signal $\boldsymbol{y}$ of dimension $D = K \cdot T$ is given by $I = \text{ceil}((D - L)/\Delta) + 1$, with Δ the window shift expressed as the number of rows in $\boldsymbol{y}$ over which the window is shifted. Δ is a multiple of K because $\boldsymbol{y}$ is a vector of concatenated frames, each with K coefficients. We denote the row indices of $\boldsymbol{W}$ and $\boldsymbol{y}$ that correspond to the coefficients in the i^{th} window by τ, with both i and τ representing natural numbers and $1 \leq i \leq I$ and $i\Delta + 1 \leq \tau \leq i\Delta + L$. For every i^{th} window a sparse representation $\boldsymbol{x}$ is computed as:

$$\min_{\boldsymbol{x}} \{ \|\boldsymbol{W}_\tau \boldsymbol{A}\boldsymbol{x} - \boldsymbol{W}_\tau \boldsymbol{y}_\tau\|_2 + \lambda \|\boldsymbol{x}\|_1 \}. \quad (9)$$

With this $\boldsymbol{x}$ the imputed spectrogram for that window is computed as $\boldsymbol{A}\boldsymbol{x}$.

The use of overlapping time windows results in multiple imputation candidates (cf. Fig. 1). In an attempt to make the imputation of each coefficient as insensitive as possible to erroneous imputations that might occur in a single window, we compute the final clean speech estimate of the d^{th} component of $\hat{\boldsymbol{y}}$, denoted by $\hat{y}_d$, as the average of all imputation candidates resulting from overlapping windows. The number of imputation candidates ranges from 1 (at the beginning and end of an utterance) to ceil(L/Δ) (in the middle).

Particularly in very noisy conditions even a window spanning 35 frames may contain zero reliable features. Clearly, in this case we cannot solve Eq. 6. Yet, despite the lack of information about the underlying signal, input must be provided to the ASR engine. Under the assumption that it is unlikely that in the presence of speech the acoustic power in all frequency bands is dominated by background noise simultaneously over a time interval of R subsequent frames, we opted to impute silence (the lowest feature values per frequency band as observed in the exemplar dictionary $\boldsymbol{A}$) in such window.

5. EXPERIMENTS

In section 4 we proposed an extension of the sparse imputation technique for discrete word recognition to continuous speech recognition. In this section we describe the experimental setup with which we tested the feasibility of this generalization on the AURORA-2 continuous digit recognition task [13]. First, we determined the maximum achievable recognition accuracy when a priori information is provided about speech and noise in the form of an oracle mask. Second, we studied the behaviour of the new imputation method using a realistic, estimated mask in the form of a harmonicity mask [14].

5.1. Recognizers and speech material

Recognition experiments were done using acoustic feature vectors consisting of Mel frequency log power spectra ($K = 23$ bands) and the recognizer described in [14, 6, 23]. In its conventional mode of operation it replaces unreliable features by estimated values using maximum likelihood per Gaussian based imputation [6]. When applying our new sparse imputation method, we used the same recognizer, but with a spectrographic mask that −after the imputation in the front-end− considers every time-frequency cell as reliable (thus requiring no additional missing data imputation). As in [7] the threshold value for computing the oracle mask was set to $10\log(\theta) = -3\,\text{dB}$, while for the harmonicity mask this value was set to $-9\,\text{dB}$. The sparse imputation itself was implemented in MATLAB. The l^1 minimization was carried out using the `SolveLasso` solver implemented as part of the `SparseLab` toolbox which can be obtained from www.sparselab.stanford.edu.

The speech material used for evaluation was taken from test set 'A' of the AURORA-2 corpus [13]. Each utterance contains a sequence of one to seven digits, artificially mixed with four different types of noise, viz. subway, car, babble, exhibition hall. We evaluated recognition accuracy as a function of SNR at the four lowest SNR levels present in the corpus, viz. $10, 5, 0$, and -5 dB. We report for the different SNR levels accuracies averaged over the four types of background noise.

5.2. Recognition experiments

Recognition performance through sparse imputation is affected by three parameters: the dictionary size N, the window size R, and the window shift Δ. In a pilot study we first established a reasonable working point for N and R. Experiments with dictionary sizes ranging from $N = 4000$ to $N = 14000$ revealed that recognition accuracy did not increase substantially with $N > 8000$. Choosing the length of the exemplars the same as in [7], viz. $R = 35$ frames, which equals the mean number of frames of a single digit, appeared to yield a reasonable balance between recognition performance and computation time. Therefore, all experiments reported in this paper used a fixed dictionary size $N = 8000$ and a fixed $R = 35$ exemplar length. In section 6.1, we examine how recognition accuracy varies when the window shift parameter is given a value of $1, 5, 10, 15, 20, 25, 30$ and 35 frames, respectively. In the remaining experiments of this paper we focus on the differences in recognition performance obtained with the two recognition approaches in more detail using only the best scoring window-shift.

6. RESULTS AND DISCUSSION

6.1. SI front-end: Window shift and recognition accuracy

Figures 2a (for the oracle mask) and 2b (for the harmonicity mask) show recognition accuracy as a function of the window shift in frames. Both figures show that recognition accuracy does not significantly decrease for window shifts up to 10 frames. For larger window shifts ($\Delta > K \cdot 10$) we can observe a trend of decreasing recognition accuracies. This is due to an increasing number of

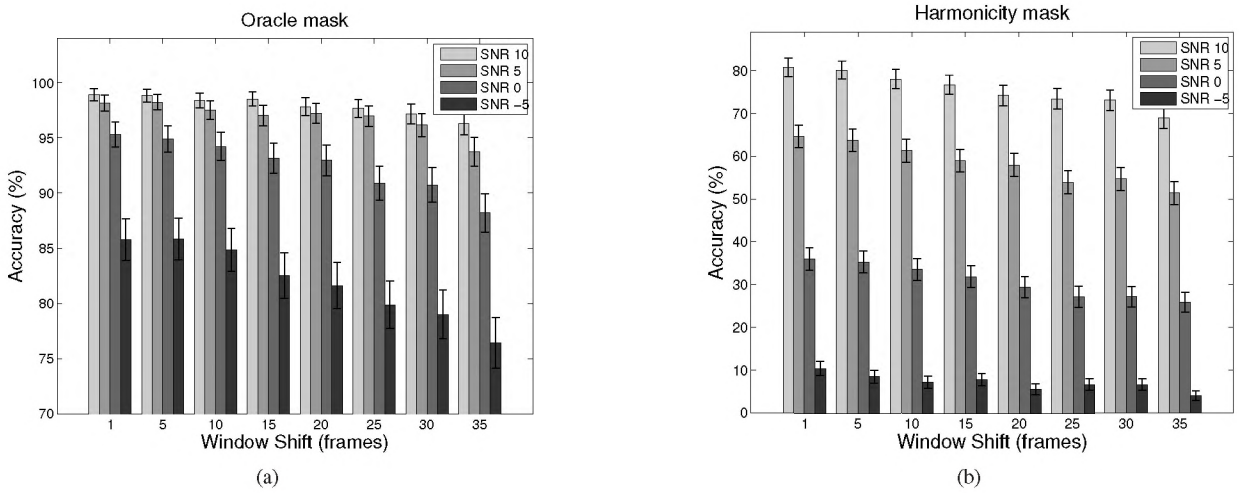

Fig. 2: AURORA-2 recognition accuracy as a function of window shift. The left panel shows results for the oracle mask and the right panel for the harmonicity mask. Window shift is expressed in frames. Note different scales of vertical axes. The vertical bars around the data points indicate 95% confidence intervals.

windows with few or no reliable features: even when using windows of 35 time frames it may occasionally happen that there are too few reliable features in the entire window for a successful data restoration. Consequently, the larger the window shift, the fewer overlapping windows there are, and the proportion of windows containing insufficient reliable features increases.

The advantage of using larger window shifts is a reduction in computational effort: using a window shift of 10 frames results in a tenfold reduction in the number of imputation candidates when compared to a one frame window shift. In the remainder of this paper we will nevertheless use a one frame window shift ($\Delta = K$), thus avoiding a possible bias from empty windows as much as possible.

6.2. SI front-end vs. CCI back-end: oracle mask

When used with an oracle mask, much higher recognition accuracies are achieved with sparse imputation than with the class conditional imputation from the baseline recognizer (cf. circles in Figure 3). In contrast with the 56% recognition accuracy obtained by the baseline decoder at SNR$= -5$ dB, 86% is a major improvement. The improvement of 30% over the baseline decoder is similar to the improvement of 31% reported in [7] for isolated digit recognition. Apparently, the fact that we used randomly selected 350 ms fragments of speech as dictionary exemplars without applying any time normalization did not severely reduce the effectiveness of the method compared to the isolated word approach.

Although the results on this connected digit recognition task cannot be directly extrapolated to an arbitrary, medium or large size vocabulary continuous speech recognition task, it is promising to observe that unreliable features can be reconstructed so well at very low SNRs, provided the reliable features can be identified correctly.

6.3. Comparing the oracle and harmonicity mask

For both imputation methods the recognition accuracies obtained with the estimated harmonicity mask (cf. diamonds in Figure 3) are much lower than with the oracle mask. Obviously, this is caused by the failure of the harmonicity mask estimation method to correctly label reliable features as such.

Mask estimation procedures may produce two kinds of errors: Unreliable features that are incorrectly labeled as reliable (false reliables) and reliable features incorrectly labeled as unreliable (false unreliables). Both types of error affect imputation: false unreliables result in fewer reliable features yielding fewer constraints that can help in finding a successful imputation, while false reliables impose incorrect constraints that may mislead the imputation. In practice one has to find a balance between these two, and generally it turns out to be most profitable to tune mask estimation routines toward less false reliables and more false unreliables. Fig. 4 shows that the percentage of features that is labeled reliable in the harmonicity mask is substantially lower than in the oracle mask, while the number of false reliables is relatively small.

Yet, the reduced number of reliable features cannot explain all observations. For instance, from Fig. 4 it can be inferred that the number of reliable features in the oracle mask at SNR $= -5$ dB is roughly equal to the number of reliable features in the harmonicity mask at SNR $= +5$ dB. Yet, the accuracies for the CCI method with harmonicity mask are higher at 5 dB than at -5 dB with oracle mask (80% vs. 55%) while for the sparse imputation the opposite is true (65% vs. 86%).

Besides the fact that the number of false reliables are bound to induce wrong imputation results which will adversely affect recognition accuracy, a third factor is at play here: The *location* of the reliable and unreliable features in the time-frequency plane. As was noted in [24], differences in recognition accuracy are difficult to explain solely in terms of the number of time-frequency cells that are considered as (un)reliable: Some incorrectly labeled spectro-temporal elements may hardly affect recognition, while others may cause the loss of information that is necessary to discriminate between different words.

Apparently, the harmonicity mask, already at moderate SNRs, fails to label such "crucial" features as reliable, making it impossible to correctly impute prior to decoding. Most likely this concerns the

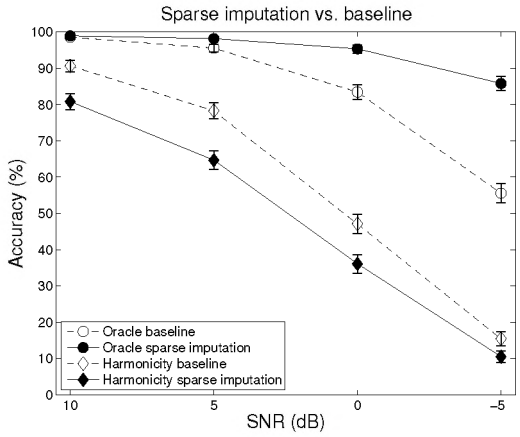

Fig. 3: Word recognition accuracy for both the baseline decoder and the sparse imputation method using the oracle mask and the harmonicity mask, respectively. The window shift is one frame. The vertical bars around the data points indicate 95% confidence intervals.

(low energy) features in the consonant parts of the utterances, which are extremely important for discriminating between different digits that have similar vowels. Given the Compressive Sensing based approach used for SI this relates to the observation in Section 3.4: Reliable features should be sufficiently randomly distributed over the time-frequency plane. We can conclude that features labeled reliable by the harmonicity mask do not cover a sufficient area in the time-frequency plane for successful imputation.

6.4. SI front-end vs. CCI back-end: harmonicity mask

An intriguing question is why the recognition accuracy with sparse imputation improves substantially (compared to the frame-based, class conditional imputation method) in all SNR conditions when an oracle mask is applied, while this is no longer true when a harmonicity mask is used.

At least two explanations spring to mind. First, in our setup, the recognizer with CCI performs *bounded* imputation, imposing the constraint that clean speech estimates cannot exceed the observed feature values. No such constraint was implemented in our SI method (yet).

A second, and probably more important difference is that the sparse imputation method is a denoising front-end approach which operates completely independently from the decoder. SI does missing data imputation on the static features only. The delta and delta-delta features offered to the decoder are computed based on these clean speech estimates. In contrast, CCI does missing data imputation on both static and derived features within the decoder. Moreover, false reliables only influence a single frame. Since with CCI the final imputation result is determined during a Viterbi search over all possible states (and thus imputation hypotheses), this gives the decoder much greater freedom in recovering from mask estimation errors. As a result, SI must be expected to be more sensitive to mask estimation errors since this method has no access to information about the fact that certain features have a different probability of being reliable in different hypotheses. This is likely to increase the risk for SI of getting a wrong imputation result, and once confronted with an incorrectly imputed speech signal, the ASR engine has no way to recover from such an error.

Table 1: Analysis of recognition errors occurring when using the harmonicity mask. We distinguish substitution (S), deletion (D), and insertion (I) errors (in %).

	SNR											
	10 dB			5 dB			0 dB			-5 dB		
Imputation	S	D	I	S	D	I	S	D	I	S	D	I
SI	8	2	9	19	3	13	40	9	16	50	23	17
CCI	4	2	4	12	2	7	35	12	7	39	41	5

The conclusion seems inevitable that what is a blessing in the case of oracle mask (good restorative power with a very limited number of reliable features) turns into a curse whenever estimated masks become too poor: Too few reliable features and/or too many false reliables cause either substitution or insertion errors. The high number of insertion errors (cf. Table 1) that occur with the sparse imputation method and a harmonicity mask confirms this.

There are several possibilities to deal with this issue. Most apparently, we have to reconsider the strategy of imputing all frames regardless of the number of reliable features. In order to suppress insertion errors, it seems wiser to only impute time-windows that contain a minimum number of reliable features. What this number should be can only be determined empirically. Such an adapted strategy would possibly reduce accuracy using the oracle mask, but most likely increase recognition accuracies using an error-prone estimated mask.

Another approach to reduce insertion errors might be to tune the mask threshold θ toward having fewer false reliables. While in the single digit experiments in [7] it was found that the used threshold of -9 dB was close to optimal, this threshold might be different for time-continuous imputation.

6.5. Extension to large vocabulary tasks

In this work, the exemplar dictionary was constructed with a random selection of fixed length exemplars from a large set of utterances. While we obtained good results using this technique for the connected digit recognition task, it seems unlikely that an exemplar dictionary containing a few thousand randomly selected items will capture all relevant variation in arbitrary speech signals. Furthermore, we obtained time-shift invariance by including exemplars extracted with a random offset. If the number of exemplars needed to capture the full variation of speech grows, so will the number of exemplars needed to ensure shift invariance.

It is not difficult to see how this process could be improved for a more general continuous speech recognition task. Shift-invariance could be handled algorithmically [25]. One possible way to prevent including too many similar exemplars while missing out on outliers would be by treating dictionary creation as a clustering problem and using the cluster means as exemplars.

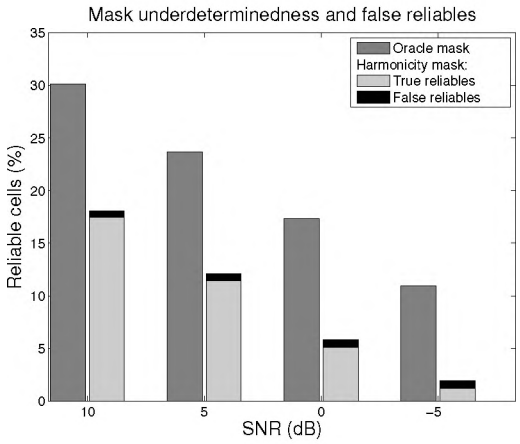

Fig. 4: Percentage of time-frequency cells classified as reliable in the oracle mask and the harmonicity mask. Additionally, the percentage of false reliables in the harmonicity mask is shown.

7. CONCLUSIONS

We described a non-parametric method based on techniques from the field of Compressive Sensing to impute noise corrupted data in ASR. The method, dubbed *sparse imputation*, builds on the assumption that the reliable features of a noise-corrupted speech unit can be represented by a sparse linear combination of clean speech exemplars from an overcomplete dictionary. Using the reliable features to determine the linear combination of exemplars, the unreliable features are replaced by the linear combination of the features from the clean speech exemplars. In previous work the dictionary was constructed with vectors that represent whole digits, limiting the applicability of the method to *isolated* digit recognition. In this paper we showed that the method can be successfully extended to the AURORA-2 *connected* digit recognition task. We applied the sparse imputation method to a sliding window, while the dictionary of exemplars was constructed from randomly selected speech fragments. In case the sliding windows are overlapping, imputation results from overlapping windows are combined by averaging.

When used with an oracle mask, the sparse imputation approach was shown to vastly outperform a frame based, class condition imputation approach at low SNRs (accuracy of 86% vs. 56% at SNR$=-5$ dB). Furthermore, we showed that we do not need maximally overlapping windows to provide robustness against windows that contain few, if any, reliable features. Our results showed that using a window shift of up to 10 frames does not result in a significant decrease in recognition accuracy while substantially decreasing the needed computational effort.

Since sparse imputation can be implemented as a denoising front-end, it potentially constitutes an elegant way to combine it with the power and efficiency of conventional continuous speech decoding. However, in order to unleash the potential shown for oracle masks in real world conditions, our experiments indicate that it is crucial for the mask estimation method to be able to label a sufficiently large proportion of features as reliable. Moreover, these features must be sufficiently spread over the time frequency plane (so that that are not only concentrated in areas that are associated with vowels) and it is important that very few false reliable errors are made. By its very nature, the sparse imputation method is easily misled when confronted with false reliables, and once an incorrect imputation result has been generated, the ASR engine has no way to recover from such errors.

Under such conditions, there is little hope that a sparse imputation front-end alone can be effective in realizing true noise robust ASR. To harness the full potential of sparse imputation in realistic conditions in a similar fashion as with oracle masks, ways must be found to either improve the mask estimation or to reduce the sensitivity of of the sparse imputation method for mask estimation errors.

8. FUTURE WORK

Earlier, we suggested that our current sliding window approach may be viewed as a first step towards applying sparse imputation to a general continuous speech recognition task. As we mentioned before, it is still an open question to what extent the proposed method is scalable to large vocabulary tasks (cf. Section 6.5). Before research into more general ASR tasks becomes opportune, however, a more urgent question needs to be answered: To what extent is it possible to increase the effectiveness of SI in combination with estimated (as opposed to oracle) missing data masks?

Currently, no estimation methods exist that can sufficiently accurately classify acoustic features as belonging to speech or background noise. This severely reduces the effectiveness of SI. Several options exist to try and avoid imputation errors due to errors in the estimated masks.

First, one might formulate additional constraints that help avoid incorrect imputation results. The CCI method applied in the baseline recognizer takes into account that in the presence of additive noise, clean speech estimates should be bounded by the observed energy, while the SI method does not. As a first order approximation, one could adapt Eq. (8) such that only unreliable values are imputed for which the clean speech estimate does not exceed the noisy feature value. Such a procedure would ensure that the final clean speech estimate is properly bounded. It will not, however, guarantee that the exemplars used to provide the clean speech estimate were the most likely set of exemplars given the extra knowledge implicitly present in the *unreliable* features. It might therefore be better to constrain the minimization itself; designing a practical implementation of such a constrained minimization and exploring its effectiveness in increasing recognition accuracy is part of our future work.

In our current set-up both the mask estimation procedure and the corresponding imputation of the missing features are implemented as a front-end which operates completely independently from the decoder. It is questionable to what extent refinements like the one suggested above will be sufficient to avoid wrong imputation results. There are various reasons why a denoising front-end is probably a suboptimal approach. First, as illustrated by Sections 6 and 7, the SI approach is vulnerable to whatever errors remain in the mask estimation procedure, and in actual practice errors will be simply unavoidable. Second, an approach with a denoising front-end does not do justice to the fact that making a distinction between foreground speech and background noise (which might also be speech) is an ill defined problem: It is virtually impossible to accurately distinguish speech from background noise on a time-frequency cell-by-cell basis without knowledge of the underlying speech. For instance, if a hissing sound is observed during the vowel /a/, it is very likely that the high frequency energy must be associated with background noise.

However if an /s/ or /f/ is produced, this same high frequency energy is very likely to constitute speech energy. Ways will need to be found to arrange a tighter coupling between the denoising front-end and the speech decoder back-end.

One possible approach would be to supplement the imputed clean speech features with a confidence measure. This might give the speech decoder more possibilities to recover from imputation errors. Applying an uncertainty decoding approach [26] seems a promising way to combine the strengths of CCI and SI. It seems reasonable to assume that such a confidence measure could be based on the number of reliable features in the window and/or the distribution of nonzero elements of the sparse representation. These confidence scores could then be presented to the decoder to indicate that the imputed acoustic observations are not really clean speech observations. To what extent SI can provide a viable framework for generating confidence scores, however, is a largely unexplored area and also needs further research.

Acknowledgments

This research was carried out in the MIDAS project, granted under the Dutch-Flemish STEVIN program. The project partners are the universities of Leuven, Nijmegen and the company Nuance. We gratefully acknowledge the many useful discussions with Lou Boves.

9. REFERENCES

[1] B. Raj and R. M. Stern, "Missing-feature approaches in speech recognition," *Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[2] M. Cooke, P. Green, L. Josifovksi, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[3] B. Raj, M. Seltzer, and R. Stern, "Reconstruction of missing features for robust speech recognition," *Speech Communication*, vol. 43, pp. 275–296, 2004.

[4] B. Raj, *Reconstruction of incomplete spectrograms for robust speech recognition*, Ph.D. thesis, Camegie Mellon University, 2000.

[5] L. Josifovski, M. Cooke, P. Green, and A. Vizinho, "State based imputation of missing data for robust speech recognition and speech enhancement," in *Proc. of Eurospeech*, 1999.

[6] H. Van hamme, "Prospect features and their application to missing data techniques for robust speech recognition," in *Proc. of INTERSPEECH-2004*, 2004, pp. 101–104.

[7] J. Gemmeke and B. Cranen, "Using sparse representations for missing data imputation in noise robust speech recognition," in *Proc. of EUSIPCO 2008*, 2008.

[8] Y. Zhang, "When is missing data recoverable?," *CAAM Technical Report TR06-15, Rice University, Houston*, 2006.

[9] J. Wright, A. Y. Yang, A. Ganesh, S. Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[10] M.N. Schmidt and R.K. Olsson, "Linear regression on sparse features for single-channel speech separation," in *Applications of Signal Processing to Audio and Acoustics, IEEE Workshop on (WASPAA)*, 2007.

[11] M. E. Davies T. Blumensath, "Compressed sensing and source separation," in *International Conference on Independent Component Analysis and Signal Separation*, sept 2007.

[12] M. V. S. Shashanka, B. Raj, P. Smaragdis, and Madhusudana V. S. Shashanka, "Sparse overcomplete decomposition for single channel speaker separation," in *Proc. of IEEE ICASSP*, 2007, pp. 641–644.

[13] H.G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. of ISCA ASR2000 Workshop, Paris, France*, 2000, pp. 181–188.

[14] H. Van hamme, "Robust speech recognition using cepstral domain missing data techniques and noisy masks," in *Proc. of IEEE ICASSP*, 2004, vol. 1, pp. 213–216.

[15] C. Cerisara, S. Demange, and J-P. Haton, "On noise masking for automatic missing data speech recognition: A survey and discussion," *Comput. Speech Lang.*, vol. 21, no. 3, pp. 443–457, 2007.

[16] A. Vizinho, P. Green, M. Cooke, and L. Josifovski, "Missing data theory, spectral subtraction and signal-to-noise estimation for robust asr: An integrated study," in *Proc. of INTERSPEECH-1999*, 1999, pp. 2407–2410.

[17] W. Kim and R. M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise," in *Proc. of IEEE ICASSP*, 2006.

[18] D. L. Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[19] E. J. Candès, "Compressive sampling," in *Proc. of the International Congress of Mathematicians*, 2006.

[20] D. L. Donoho, "For most large underdetermined systems of linear equations the minimal l1-norm solution is also the sparsest solution," *Communications on Pure and Applied Mathematics*, vol. 59, no. 6, pp. 797–829, 2006.

[21] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[22] J. Gemmeke and B. Cranen, "Sparse imputation for noise robust speech recognition using soft masks," in *Accepted at ICASSP 2009*, 2009.

[23] H. Van hamme, "Handling time-derivative features in a missing data framework for robust automatic speech recognition," in *Proc. of IEEE ICASSP*, 2006.

[24] M. Seltzer, B. Raj, and R. Stern, "A bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, vol. 43, pp. 379–393, 2004.

[25] M. Morup and M. N. Schmidt, "Shift invariant sparse coding of image and music data," *Submitted to Neural Networks*, 2008.

[26] V. Stouten, H. Van hamme, and P. Wambacq, "Model-based feature enhancement with uncertainty decoding for noise robust asr," *Speech Communication*, vol. 48, no. 11, pp. 1502–1514, nov 2006.